

Reply trees in Twitter: data analysis and branching process models

Ryosuke Nishi^{1,2,6} · Taro Takaguchi^{1,2,7} · Keigo Oka^{2,3} · Takanori Maehara^{1,2,8} · Masashi Toyoda⁴ · Ken-ichi Kawarabayashi^{1,2} · Naoki Masuda⁵

Received: 29 June 2015 / Revised: 22 April 2016 / Accepted: 24 April 2016 / Published online: 10 May 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract Structure of networks constructed from mentioning relationships between posts in online media may be valuable for understanding how information and opinions spread in these media. We crawled Twitter to collect tweets and replies to construct a large number of so-called reply trees, each of which was rooted at a tweet and joined by replies. Consistent with the previous literature, we found that the empirical trees were characterized by some long path-like reply trees, large star-like trees, and long irregular trees, although their frequencies were not high. We tested several branching process models to explain the empirical

frequency of these types of reply trees as well as more basic quantities such as the distributions of the size and depth of the reply tree. Based on our modeling results, we suggest that the in-degree of the tweet that initiates a reply tree (i.e., the number of times that the tweet is directly mentioned by other reply posts) may play an important role in forming the global shape of the reply tree.

Keywords Reply tree · Twitter · Branching process · Data analysis

✉ Taro Takaguchi
t_takaguchi@nii.ac.jp

✉ Naoki Masuda
naoki.masuda@bristol.ac.uk

¹ National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

² JST, ERATO, Kawarabayashi Large Graph Project, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

³ Department of Computer Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

⁴ Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan

⁵ Department of Engineering Mathematics, University of Bristol, Woodland Road, Clifton, Bristol BS8 1UB, UK

⁶ Present Address: Department of Mechanical and Aerospace Engineering, Graduate School of Engineering, Tottori University, 4-101 Minami, Koyama, Tottori 680-8552, Japan

⁷ Present Address: National Institute of Information and Communications Technology, 4-2-1 Nukui-Kitamachi, Koganei, Tokyo 184-8795, Japan

⁸ Present Address: Department of Mathematical and Systems Engineering, Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka 432-8561, Japan

1 Introduction

Information spreading plays a fundamental role in triggering collective actions in human society on a large scale. A classical example is diffusion of technological innovation, in which individuals receiving information on a new technology from other peers may decide to adopt the technology (Rogers 2003; Easley and Kleinberg 2010). Other examples include fads (Gladwell 2000), social mobilization (Lotan et al. 2011; Banõs et al. 2013; Conover et al. 2013), marketing (Leskovec et al. 2007a; Easley and Kleinberg 2010), voter turnout (Bond et al. 2012), responses to natural disasters (Sano et al. 2013; Sasahara et al. 2013), and circulation of new scientific publications (Thelwall et al. 2013) to name but a few.

Network analysis has been a useful tool for understanding information spreading both online and offline. In particular, owing to increasing amounts of users' activity and availability of data, various online social media ranging from micro-blogging services (e.g., Twitter), to social networking services (e.g., Facebook) have been analyzed as networks. In networks of users, a node represents a user, and a link represents a relatively static dyadic relationship between two users such as

followership in Twitter and mutual friendship in Facebook. An alternative construct, which we focus on in the present study, is networks of posts; a node represents a post by a user, and a link represents a reference relationship from a post to a previous post. Such a network is usually tree-like with possible branching and without confluence. The reference relationship implies that a post spreads information relayed from a previous post. Therefore, a network of posts is considered to be a direct derivative of information spreading. It should be noted that, in contrast to the case of user networks, a user may appear as different nodes in a network of posts. Networks of posts have been studied in Twitter (Kumar et al. 2010; Kwak et al. 2010; Bakshy et al. 2011; Cogan et al. 2012), Facebook (Sun et al. 2009; Cheng et al. 2014), blogs (Leskovec et al. 2007b; McGlohon et al. 2007; Götz et al. 2009), Flickr (Cha et al. 2009), discussion threads (Gómez et al. 2008; Kumar et al. 2010; Gómez et al. 2011; Wang et al. 2012; Gómez et al. 2013), and e-mail (Liben-Nowell and Kleinberg 2008; Golub and Jackson 2010; Wang et al. 2011).

In a network of posts, an initial post located at the root of the network may induce a cascade of responses of different magnitudes and spatiotemporal patterns. Structure of such a network seems to inform us of the nature of the cascade (Leskovec et al. 2007b; Iribarren and Moro 2009; Liben-Nowell and Kleinberg 2008; Cha et al. 2010; Golub and Jackson 2010; Kumar et al. 2010; Iribarren and Moro 2011; Wang et al. 2011; Wang et al. 2012). For example, the size of the cascade defined by the number of nodes in the network is a simple measure of the extent to which the initial post has involved other users. In addition, networks of the same size may have different shapes. An initial post may diffuse by forming a long chain-like network to eventually involve 100 other posts. A different initial post may receive 100 direct replies, and then, the cascade may terminate without further diffusion, resulting in a star network. Although the size of the cascade is the same in the two cases, the way information is communicated during the cascade may be different. A long chain-like network may be formed by alternately replying behavior between two users with which the two users end up detailing the topic; a star network does not allow this interpretation (Cogan et al. 2012). The structure of networks of posts may also tell us the importance of individual users and posts involved in information cascades (Cha et al. 2010; Kwak et al. 2010; Weng et al. 2010; Bakshy et al. 2011; Wang et al. 2011; Banõs et al. 2013). The previous studies used structural information obtained from networks of posts for practical applications. Examples include classification of topics without text mining (McGlohon et al. 2007; Kumar et al. 2010; Gómez et al. 2011), quantification of how controversial a post is in online discussion threads (Gómez et al. 2008), and predictions of the final size of an information cascade (Cheng et al. 2014).

In the present study, we analyze a large data set of trees formed by posts in Twitter, which we call reply trees. We operationally distinguish three types of post in Twitter in the present paper: tweet, reply, and retweet. By convention, we do not include reply and retweet to the definition of tweet. Then, we construct trees from the tweets and replies that we have collected. In short, a reply tree is rooted at a tweet and involves replies that refer to the tweet directly or indirectly. We analyze structural properties of empirical reply trees and propose branching process models for them.

We use Twitter data because Twitter is suitable for studying diffusion processes for several reasons (Kwak et al. 2010; Bakshy et al. 2011; Bollen et al. 2011; Dodds et al. 2011; Lotan et al. 2011; Bliss et al. 2012; Cogan et al. 2012; Banõs et al. 2013; Conover et al. 2013; Sasahara et al. 2013). First, Twitter is devoted to information diffusion. This situation contrasts with that for other media such as Facebook in which mutual endorsement is more emphasized. Second, Twitter users communicate in standardized ways. Tweets are restricted to 140 characters, and retweets and replies, which have to follow a given standardized format, are the only modes allowed with which users can directly respond to previous posts. Third, Twitter data can be collected on a large scale with the use of the application programming interface (API).

A majority of previous literature on networks of posts in Twitter seems to have focused on networks of retweets (e.g., Kwak et al. 2010; Bakshy et al. 2011) rather than those of replies (but see Kumar et al. 2010; Cogan et al. 2012). However, we focus on replies in the present study for two reasons. First, replies are considered to be more informative about the relationships between users than retweets are (Sousa et al. 2010; Gonçalves et al. 2011; Bliss et al. 2012). Second, replies are suggested to convey emotional responses of users (Dodds et al. 2011; Bliss et al. 2012), and collective emotions and moods in Twitter often covary with the results of collective actions, presumably induced by information spreading, such as dynamics of stock prices (Bollen et al. 2011).

2 Methods

2.1 Data

We collected mentioning relationships (i.e., one post mentioning another post) between pairs of public posts in Twitter from the December 1–9, 2011, using Twitter API as follows. First, on March 15, 2011, we manually selected 26 Japanese celebrity users with many followers as seed users. Second, we collected the posts, i.e., tweets (excluding replies and retweets by definition), replies, and retweets, made by the seed users using the user timeline API provided by Twitter. We collected all the posts made by the seed users between March 15 and December 9,

2011. Between the 11th and 15th March, we collected 3000 most recent posts for each user due to the limitation imposed by Twitter on the user timeline API. Third, we added about 1000 users who received the largest number of responses within the data collected up to the previous step. Here, a response to a user X is operationally defined as either a retweet to an X 's post containing at least one Japanese character or a reply containing a Japanese character to an X 's post. Fourth, we collected posts made by each of the newly added users. We collected all the posts of each newly added user between the time at which the newly added user was detected for the first time and December 9, 2011. We also collected the most recent 3000 posts of the newly added user before the time the user was detected for the first time, under the condition that the posts are dated March 11, 2011 or later. Fifth, we repeated the third and fourth steps a large number of times to expand the set of the users and posts.

We excluded the replies and retweets that did not explicitly contain the IDs of the posts that these replies and retweets referred to, because construction of directed links was not straightforward for these replies and retweets. As a result, restricted to the period between December 1–9, 2011, we obtained 505,557 users and 57,982,740 posts including 24,280,912 replies and 5,478,846 retweets. Before analyzing the data, we anonymized the user IDs and discarded the contents of the posts while keeping the information about the mentioning relationships between all pairs of posts. We discarded retweets and defined tweets and replies as nodes. A directed link is defined as dyadic relationship from the mentioning post, which is a reply (because we have discarded retweets), to the mentioned post, which is either a tweet or reply.

2.2 Reply tree

A reply tree is defined as a directed tree composed of a tweet, which is located at the root of the tree, and all replies from which the tweet is reached along directed paths (Fig. 1). By definition, a directed tree is a directed network in which any node is connected to the root by a unique directed path to the root. It is a tree if the direction of the link is neglected. We refer to directed tree as tree. Other studies also investigated reply trees in Twitter, although the definition of node and link may be slightly different from ours (Kumar et al. 2010; Cogan et al. 2012). The out-degree is the number of mentioning that a node has made. The out-degrees of a tweet (i.e., root node) and a reply (i.e., nonroot node) are equal to zero and one, respectively. The in-degree of a node, denoted by k , is the number of replies that the node has received.

Unless otherwise stated, we exclude isolated tweets, i.e., those never mentioned by any reply post within the

observation period, from the definition of the reply tree. Therefore, the size of a reply tree in terms of the number of nodes, denoted by S , is at least two. Owing to our data collection method, we exhaustively collected all reply trees containing at least one sampled user unless the tweet at the root of the reply tree occurred before the observation period. We discarded reply trees whose root (i.e., tweet) was dated before the observation period. Then, there are 2,170,021 reply trees, which are by definition as many as the tweets that are posted in the observation period and have been mentioned at least once. The number of replies summed over all the reply trees is equal to 6,903,147.

We cannot exclude the possibility that a reply tree grows after the observation period by receiving a new reply. If a reply tree starts from a tweet located near the end of the observation period, the tree is likely to grow even after the observation period. We confirmed that statistics of reply trees calculated from the entire data set did not considerably differ from those calculated from the partial data set composed of the reply trees whose roots were located in the first half of the observation period (Appendix 1). We focus on the entire data set in the following.

3 Results of data analysis

The size of reply tree (Kumar et al. 2010; Wang et al. 2011, 2012; Gómez et al. 2013) (also called cardinality Cogan et al. 2012), denoted by S , is equal to the number of nodes in a reply tree. For example, the reply tree shown in Fig. 1 has $S = 12$. The survivor function (also called complementary cumulative distribution) of S defined by $P_{\text{surv}}(S) \equiv \sum_{S'=S}^{\infty} P(S')$, where $P(S)$ represents the frequency of trees of size S , is shown by the solid line in Fig. 2a. Hereafter, P_{surv} and P represent a survivor function and frequency distribution, respectively. The tail part of the survivor function is roughly approximated by a power law $P_{\text{surv}}(S) \propto S^{-3}$, which implies $P(S) \propto S^{-4}$. Because the power-law exponent for $P(S)$ is larger than three, we conclude that the distribution of S is not long-tailed. Consistent with this claim, the coefficient of variation (CV), i.e., standard deviation divided by the mean, of S is equal to 0.89. This value is not considered to be large; the CV is equal to unity for the exponential distribution and considerably larger than unity for a long-tailed distribution. The short-tailed nature of $P(S)$ qualitatively agrees with some previous results (Leskovec et al. 2006; Wang et al. 2011) and different from others (Leskovec et al. 2007b; Götz et al. 2009; Kumar et al. 2010; Li et al. 2012).

The depth of reply tree, denoted by D , is defined as the maximal distance from the root (Kumar et al. 2010; Wang et al. 2011; Gómez et al. 2013). For example, the reply tree shown in Fig. 1 has $D = 5$. The survivor function of D is

Fig. 1 Schematic of a reply tree. Root node 1 is a tweet. The other 11 nonroot nodes are replies. This reply tree has size $S = 12$ and depth $D = 5$. The root has in-degree $k_t = 2$. For segment $\{2, 5, 7, 9\}$, node 2 is the start node with in-degree $k_s = 2$, node 9 is the end node with in-degree $k_e = 3$, and the length is equal to $\lambda = 3$

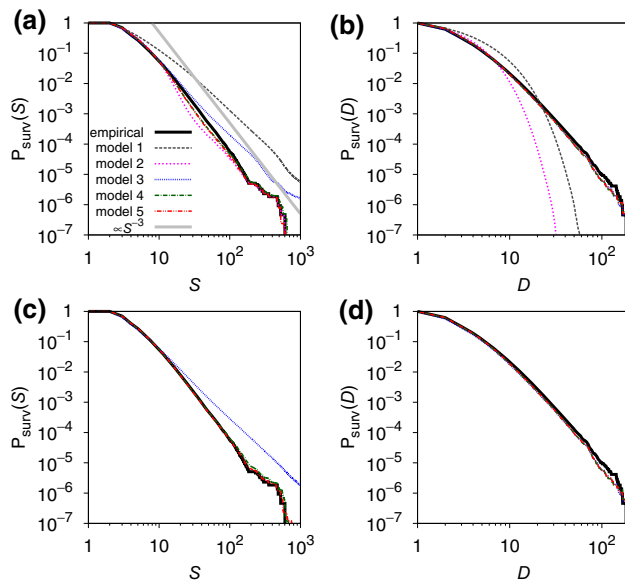
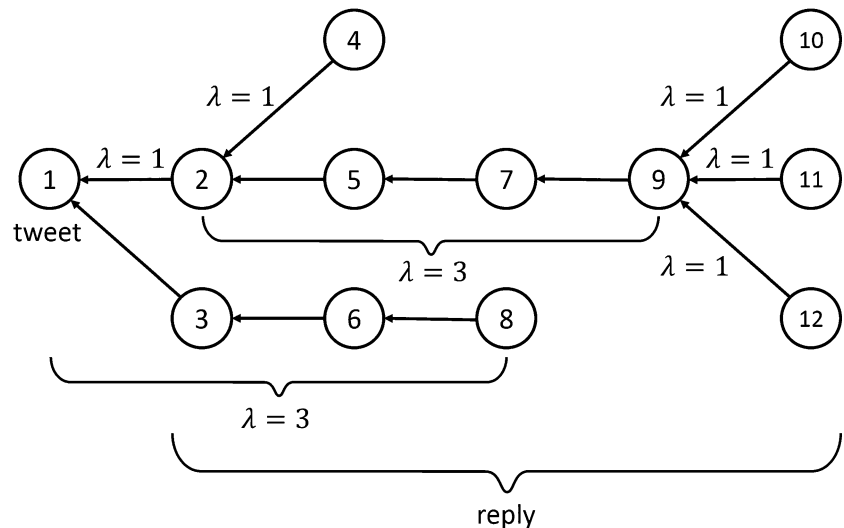


Fig. 2 Distributions of the size, S , and depth, D , of reply trees. **a** Survivor function of S (i.e., probability that the size is at least S) with $\rho = 0$. **b** Survivor function of D with $\rho = 0$. **c** Survivor function of S with $\rho = 0.7$. **d** Survivor function of D with $\rho = 0.7$

shown by the solid line in Fig. 2b. The distribution of D is short-tailed with a CV value of 1.01, which is consistent with previous results (Kumar et al. 2010).

The distributions of S and D are not independent. The joint distribution, denoted by $P(S, D)$, is shown in Fig. 3a as a heat map and looks similar to that reported in a previous study (Cogan et al. 2012). As informally classified in Cogan et al. (2012), there are at least three characteristic types of reply tree. First, some reply trees are close to long paths having $S \approx D + 1$ (note that $D + 1$ is the minimum possible value of S) and large S (equivalently, large D). In fact, some points are located near the diagonal for large

S in Fig. 3a. A long path-like reply tree has just one long branch, which involves most of the nodes in the reply tree. It is equivalent to a skinny tree referred to in Kumar et al. (2010). Second, there exists a portion of large star-like reply trees defined by $D \approx 1$ and large S . In typical large star-like trees, the tweet directly receives many replies, whereas these replies receive few other replies. Third, so-called large irregular trees are defined by large S and intermediate values of D .

It should be noted that the results shown in Fig. 3a are not comparable with those in Gómez et al. (2013), which has also investigated the relationship between S and D . This is because D values averaged over discussion trees possessing the same value of S are examined in Gómez et al. (2013). In contrast, we are concerned with distributions of S and D for individual trees.

The survivor function of the in-degree (i.e., k) is shown by the thin solid line in Fig. 4. In this figure, we included the nodes with $k = 0$ when calculating the survivor function. By definition, $P_{\text{surv}}(k)$ at $k = 1$ is the fraction of nodes having in-degree at least one. The remainder of the nodes has $k = 0$. In the figure, we also show the survivor function of the in-degree of tweet (i.e., $P_{\text{surv}}(k_t)$, where k_t is the in-degree of tweet; t for tweet; shown by the dashed line) and that of reply (i.e., $P_{\text{surv}}(k_r)$, where k_r is the in-degree of reply; r for reply; shown by the dotted line). The figure indicates that the tails of $P_{\text{surv}}(k)$ and $P_{\text{surv}}(k_r)$ roughly obey $\propto k^{-2.3}$ (thick solid line). The fact that $P_{\text{surv}}(k)$, $P_{\text{surv}}(k_t) \propto k^{-2.3}$ is translated into $P(k)$, $P(k_t) \propto k^{-3.3}$ implies that the distributions are not long-tailed. The distribution of k_r has a shorter tail. Consistent with these results, the CV values for k , k_t , and k_r calculated exclusive of the nodes with degree zero are not large and equal to 0.94, 1.28, and 0.19, respectively. The CV values with the zero-degree nodes included are equal to 3.18, 5.77, and 0.94 for

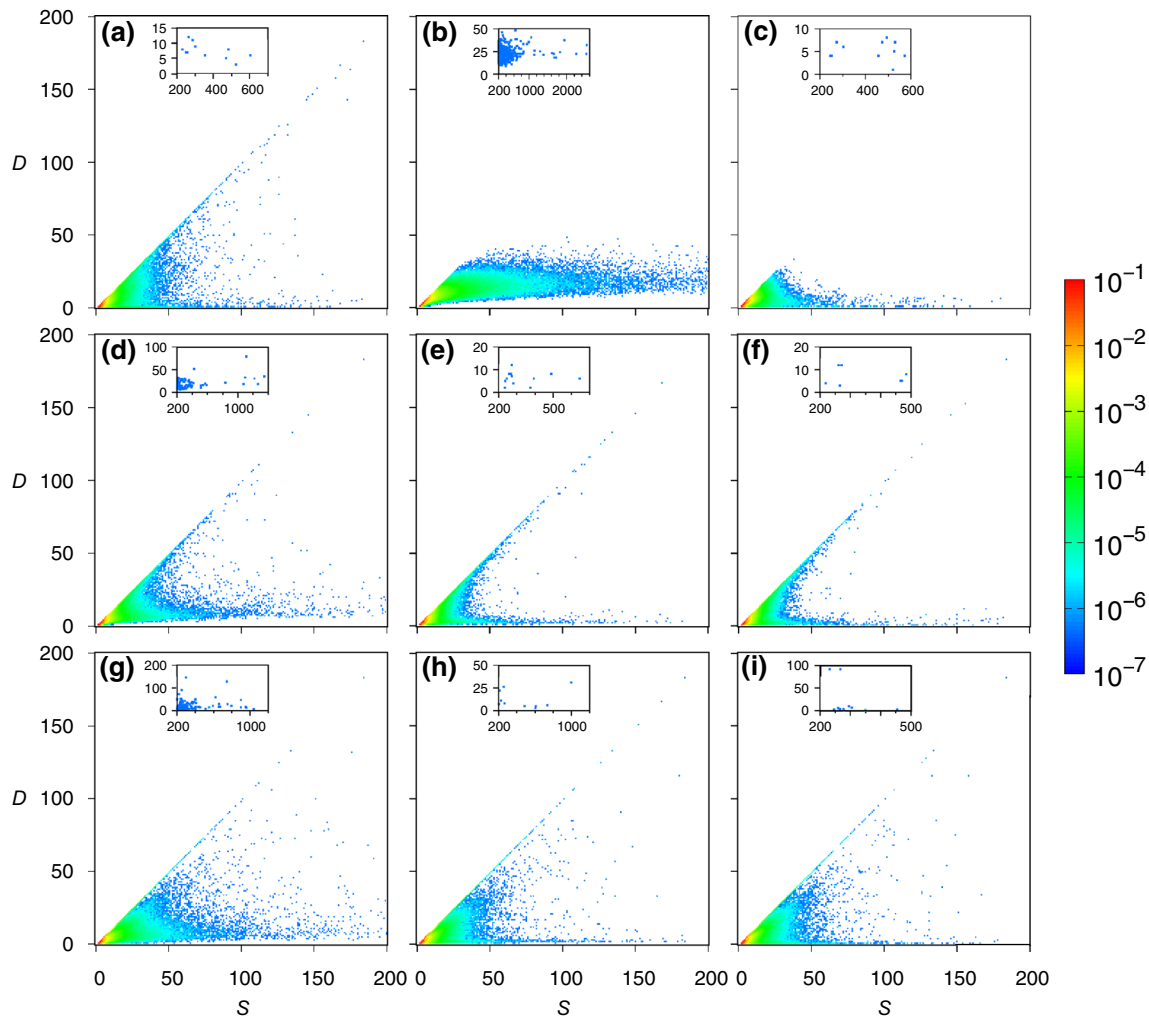


Fig. 3 Joint distribution of size and depth of reply trees. **a** Empirical data. **b** Galton–Watson process (model 1). **c** Correlated Galton–Watson process (model 2). **d** Model 3. **e** Model 4. **f** Model 5. **g** Model

3 with copula. **h** Model 4 with copula. **i** Model 5 with copula. We set $\rho = 0$ in (**b–f**) and $\rho = 0.7$ in (**g–i**). The insets show the results for large S

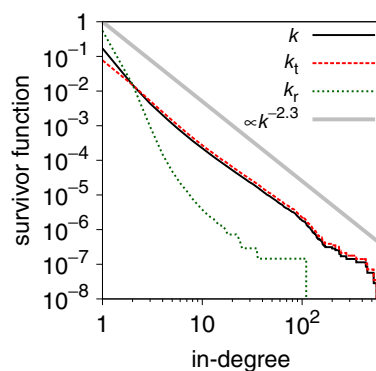


Fig. 4 Survivor functions of the in-degree (i.e., probability that the in-degree is at least k) for all nodes, $P_{\text{surv}}(k)$ (thin solid line), tweets, $P_{\text{surv}}(k_t)$ (dashed line), and replies, $P_{\text{surv}}(k_r)$ (dotted line). The thick solid line, representing $\propto k^{-2.3}$, is a guide to the eyes

k , k_t , and k_r , respectively. The CV values for k and k_t in this case are large because, in general, the CV is large when the mean value is very small. In the present case, the mean degrees are much smaller with the inclusion of zero-degree nodes (i.e., $\langle k \rangle = 0.20$, $\langle k_t \rangle = 0.11$, and $\langle k_r \rangle = 0.57$, where $\langle \cdot \rangle$ denotes the mean) than without them (i.e., $\langle k \rangle = 1.16$, $\langle k_t \rangle = 1.38$, and $\langle k_r \rangle = 1.03$). The observation that the in-degree distribution for the roots (i.e., tweet) is longer-tailed than that for the nonroots (i.e., replies) is consistent with the results for conversation trees in USENET (Kumar et al. 2010) and trees composed of email correspondences (Wang et al. 2011). Also see Wang et al. (2011) and Gómez et al. (2013) for a model incorporating this factor.

Next, we look at so-called segments, which are defined as maximal chains without branching (called path subgraph

in Lovejoy and Loch 2003). The length of the segment, denoted by λ , is equal to the number of links in the segment. By definition, a segment of length λ is composed of a linearly connected $\lambda + 1$ nodes such that the $\lambda - 1$ nodes in the middle, which are necessarily replies, have in-degree one. We refer to the node that occurs the earliest in time in a segment as the start node of the segment. A start node is either a reply at which branching occurs (i.e., $k \geq 2$) or a tweet (i.e., $k \geq 1$). The node occurring the latest in time in a segment is referred to as the end node of the segment. The in-degree of an end node is either zero or at least two. In the former case, the end node is a leaf of the reply tree. In the latter case, the end node is the start node of other segments. In the reply tree shown in Fig. 1, the two segments starting from node 2 have lengths $\lambda = 1$ and $\lambda = 3$. They end at nodes 4 and 9, respectively.

The survivor function of the length of segment, $P_{\text{surv}}(\lambda)$, is shown by the solid line in Fig. 5. The mean of λ is equal to 2.15. The distribution is roughly approximated by a log-normal distribution $P(\lambda) = \exp\left\{-[\ln(\lambda - 1) - \mu]^2 / 2\sigma^2\right\} / [\sqrt{2\pi}\sigma(\lambda - 1)]$ with $\mu = 0$ and $\sigma = 1$ (dashed line showing the survivor function of the fitted log-normal distribution). The empirical distribution $P(\lambda)$ is not long-tailed, with a CV value of 0.97. Although there are segments whose λ is much larger than the mean, their frequency is too small to qualify the distribution to be long-tailed.

4 Modeling with branching processes

In this section, we investigate models of reply trees. The goal of the modeling is to approximate properties of the empirical data shown in Sect. 3 to illuminate generative mechanisms of reply trees. It should be noted that our aim

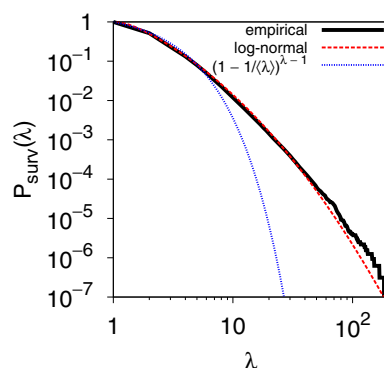


Fig. 5 Survivor function of the length of segment (solid line). The log-normal distribution with $\mu = 0$ and $\sigma = 1$ is shown by the dashed line as a guide to the eyes. The geometrical distribution whose parameter p is estimated from the empirical data is shown by the dotted line

is not to generate synthetic reply trees for certain tasks, but to understand the mechanisms governing the growth of empirical reply trees.

The models introduced in the following are variants of branching processes (Harris 1963; Kimmel and Axelrod 2002). Branching processes were employed in previous literature for modeling information spreading online (Fortunato and Castellano 2007; Vazquez et al. 2007; Liben-Nowell and Kleinberg 2008; Iribarren and Moro 2009; Golub and Jackson 2010; Kumar et al. 2010; Wang et al. 2011; Iribarren and Moro 2011; Li et al. 2012; Gómez et al. 2013; Jo et al. 2014; Gleeson et al. 2014). We build the models by combining empirical distributions related to k and λ in different ways. The assumptions underlying each model, i.e., unconditional and conditional probability distributions used in the model, are summarized in Table 1. In words, we consider a simple Galton–Watson process in Sect. 4.2.1 and its extension with degree–degree correlation in Sect. 4.2.2 to find that they do not produce the structural properties of the empirical reply trees. Then, to improve fitting of the model to the empirical data, we incorporate into the model the distribution of segment length λ in Sect. 4.3.1 and the correlation between λ and the in-degree of the tweet of the tree, i.e., k_t in Sect. 4.3.2. Finally, we take into account the correlation between k_t and the in-degree of the end node of segments, i.e., k_e in Sect. 4.3.3.

It is worth noting at this point how reply trees in Twitter were previously modeled. In Kumar et al. (2010), three models were considered. In the first model, a new reply chooses which tweet or reply to attach to with the probability proportional to a linear combination of its in-degree and age. The second model extends the first model by considering the authorship of each reply. The third model is a branching process model with multiple types of replies each of which is associated with a separate in-degree distribution. The type of each reply is estimated by the expectation–maximization algorithm. It is difficult to conclude which of the three models fits to Twitter reply trees, because the main focus of Kumar et al. (2010) was on data different from those obtained from Twitter and the

Table 1 Empirical distributions used in each model. For models 3, 4, and 5, copula variants were also examined in Sect. 4.4.

Model	Distributions used
Galton–Watson (model 1)	$P(k)$
Correlated Galton–Watson (model 2)	$P(k_t)$, $P(k k_{\text{prev}})$
Model 3	$P(k_t)$, $P(\lambda)$, $P(k_e)$
Model 4	$P(k_t)$, $P(\lambda k_t)$, $P(k_e)$
Model 5	$P(k_t)$, $P(\lambda k_t)$, $P(k_e k_t)$

k : in-degree, k_t : in-degree of a tweet; k_{prev} : in-degree of the previous node; λ : length of a segment; k_e : in-degree of the end node of a segment

comparison between the model results and the Twitter data was not provided in quantitative terms.

4.1 Measurements

For each model, we generate the joint distribution of the size and depth of reply tree, $P(S, D)$, using the same number of samples as that for the empirical data (i.e., $N = 2,170,021$), and compare it with the empirical distribution shown in Fig. 3a.

For quantitative comparisons, we also generate $N = 5 \times 10^7$ synthetic reply trees from each model and carry out the following analysis. First, we measure marginalized survivor functions of the size and depth of reply trees, i.e., $P_{\text{surv}}(S)$ and $P_{\text{surv}}(D)$. Second, we measure the fraction of long path-like reply trees, large star-like reply trees, and large irregular reply trees as follows. We define the long path-like tree as a reply tree satisfying $S - d_1 \leq D \leq S - 1$ and $S \geq 50$, where d_1 is a threshold value and presumably much smaller than D . Only exact paths are counted if $d_1 = 1$. Similarly, the large star-like tree is defined by $1 \leq D \leq d_2$ and $S \geq 50$, where d_2 is a presumably small threshold. The large irregular tree is defined by $d_3 \leq D \leq S - d_4$ and $S \geq 50$, where d_3 and d_4 are thresholds. We measure the fraction of long path-like trees, that of large star-like trees, and that of large irregular trees, relative to all generated reply trees for various threshold values.

4.2 Galton–Watson process and its correlated variant

4.2.1 Galton–Watson process (model 1)

Given the moderately heterogeneous in-degree distribution of the reply trees, the simplest model is probably the Galton–Watson branching process, in which we draw the in-degree of each node from the empirical degree distribution $P(k)$ (Harris 1963; Kimmel and Axelrod 2002). The Galton–Watson process defines model 1 (Table 1). Because it always holds that $S \geq 2$ according to our convention, we discard samples that have yielded an isolated root node, which would result in $S = 1$.

The distributions of S and D produced by model 1 are compared with the empirical distribution in Fig. 2a, b, respectively. The model overestimates the probability at large S and underestimates the probability at large D .

The joint distribution of S and D obtained from model 1 is shown in Fig. 3b. We observe that the model does not generate long path-like reply trees (i.e., near the diagonal for large S and D), which contrasts with the empirical data (Fig. 3a). This result is consistent with Fig. 2b, which indicates the lack of trees with large D for model 1. More quantitatively, the fraction of long path-like trees as defined in Sect. 4.1 is almost equal to zero for the range of d_1 shown in Fig. 6a. Long path-like trees are absent because the length of segment, λ , for model 1 (i.e., Galton–Watson process) obeys the geometric distribution

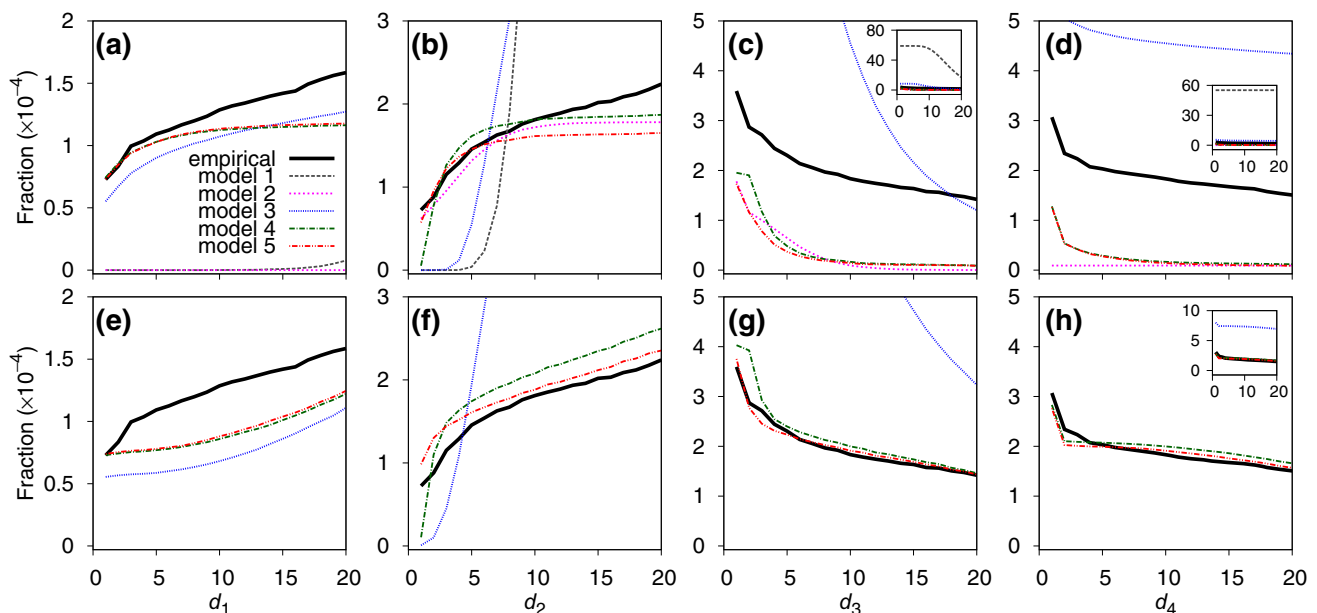


Fig. 6 Fraction of three types of reply trees. The results for the five models with $\rho = 0$ are shown in (a–d). The results for models 3, 4, and 5 with copula with $\rho = 0.7$ are shown in (e–h). **a** Fraction of long path-like trees with $\rho = 0$. **b** Fraction of large star-like trees with $\rho = 0$. **c** Fraction of large irregular trees with $\rho = 0$ and $d_4 = 10$. **d**

Fraction of large irregular trees with $\rho = 0$ and $d_3 = 10$. **e** Fraction of long path-like trees with $\rho = 0.7$. **f** Fraction of large star-like trees with $\rho = 0.7$. **g** Fraction of large irregular trees with $\rho = 0.7$ and $d_4 = 10$. **h** Fraction of large irregular trees with $\rho = 0.7$ and $d_3 = 10$. The insets show the results for large fraction values

$P(\lambda) = (1-p)p^{\lambda-1}$, where $p = P(k=1)$. The geometric distribution with the value of p estimated from the empirical data is shown by the dotted line in Fig. 5, confirming that model 1 does not produce long path-like trees as observed in the empirical data. The CV for λ obtained from model 1 is equal to 0.77, which is considerably smaller than that for the empirical data, i.e., 0.97 (Sect. 3). Model 1 does not produce a realistic fraction of large star-like reply trees, either (see Fig. 6b where d_2 is small). Finally, model 1 overestimates the frequency of large irregular trees relative to the empirical data (insets of Fig. 6c, d). In summary, the standard Galton–Watson process does not reproduce chief statistical characteristics of reply trees observed in the empirical data.

4.2.2 Correlated Galton–Watson process (model 2)

In an attempt to improve fitting of the model to the empirical data, we consider the so-called correlated Galton–Watson process (model 2). In this model, the in-degree of replies is drawn from conditional distribution $P(k|k_{\text{prev}})$, where k_{prev} is the degree of the previous node (defined as the node that the focal reply node mentions). By convention, $P(X|Y)$ here and in the following indicates the distribution of X conditioned by the value of Y . The correlated Galton–Watson process is a special case of the so-called macro process model (Olofsson 1996). In fact, Fig. 7a indicates that the in-degree of a node considerably decreases on an average as the in-degree of the previous node increases, which is consistent with the assumption of model 2. To initiate a reply tree, we draw the in-degree of tweet, $k_t (\geq 1)$, from the empirical distribution of the in-degree constructed from all tweets with $k_t \geq 1$, i.e., $P(k_t)$, because a tweet does not have a previous node.

Figure 2a indicates that model 2 produces a distribution of S similar to the empirical one despite some noticeable deviation in a middle range of S . Figure 2b indicates that model 2 underestimates the probability of D at large values of D relative to the empirical data. Figure 6a, together with the joint distribution $P(S, D)$ shown in Fig. 3c, indicates that model 2 does not produce long path-like reply trees, similarly to model 1. The CV value for λ obtained from model 2 is equal to 0.79, which is close to the value for model 1 and smaller than that for the empirical data. This result is consistent with the fact that model 2 produces the geometrical distribution of λ , i.e., $P(\lambda) = (1-p)p^{\lambda-1}$, where $p = P(k=1|k_{\text{prev}}=1)$. Model 2 produces a realistic frequency of large star-like reply trees across a range of threshold d_2 (Fig. 6b). However, model 2 by far underestimates the frequency of large irregular trees in an entire range of d_3 and d_4 (Fig. 6c, d).

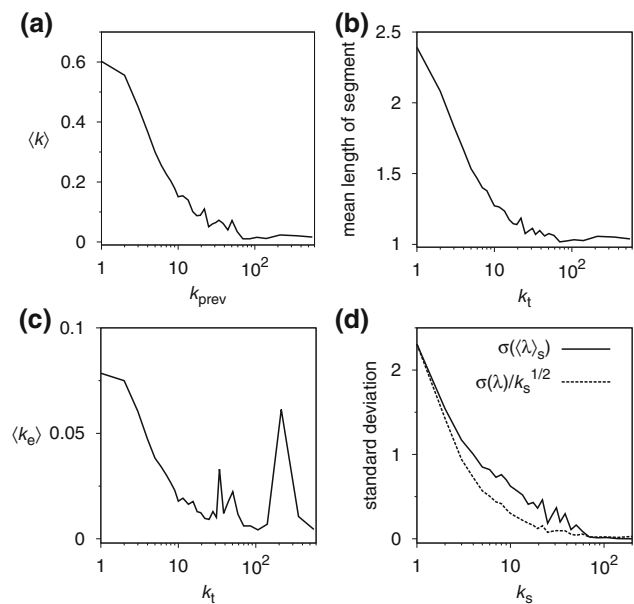


Fig. 7 Further statistical properties of empirical reply trees. **a** Mean in-degree of the node plotted against the in-degree of the previous node, k_{prev} . **b** Mean length of segment plotted against the in-degree of the tweet at the root of the reply tree. **c** Mean in-degree of the end node of a segment plotted against the in-degree of the tweet at the root of the reply tree. **d** Standard deviation of $\langle \lambda \rangle_s$ plotted against k_s (solid line). The standard deviation of λ divided by $\sqrt{k_s}$ is shown by the dotted line for comparison. In each panel, we smoothed the plots to reduce fluctuations due to the shortage of samples with a large in-degree. To be precise, we generated the data points at large in-degree values by partitioning the horizontal axis into windows and pooling instances within each window

4.3 Models that explicitly use the empirical distribution of the segment length

At best, the Galton–Watson processes (models 1 and 2) produce a realistic fraction of large star-like reply trees but not long path-like trees or large irregular trees. The models do not produce realistic distributions of S and D , either. Therefore, we explore models that go beyond the family of conventional branching process. In the models considered in this and the following sections, we draw λ from empirically determined distributions. In fact, segments are generated by users' microscopic behavior. We have decided not to model this factor, and the limitation of the present approach will be discussed in Sect. 5.

4.3.1 Model 3

We extend the Galton–Watson process to define model 3 as follows. First, we draw the in-degree of the tweet, k_t , from the empirical distribution $P(k_t)$, as is done in model 2. Second, we draw the length of each of the k_t segments starting from the root independently from the empirical

distribution, $P(\lambda)$. Third, for each segment, the in-degree of the end node, denoted by k_e , is drawn from $P(k_e)$, which is constructed from all end nodes of segments in all empirical reply trees. The use of different in-degree distributions for tweets and replies is motivated by a clear difference between $P(k_t)$ and $P(k_r)$ in the empirical data shown in Fig. 4. It should be noted that $P(k_e = k') \propto P(k_r = k')$ for $k' \geq 2$. It should be also noted that, because the end node of a segment is either a leaf or a branching node, $P(k_e = 1) = 0$. Fourth, if the end node of a segment attains $k_e \geq 2$, the lengths of k_e segments starting from this node are independently drawn from $P(\lambda)$. We repeat the procedure until all branches terminate.

The joint distribution of S and D obtained from model 3 is shown in Fig. 3d. The model produces some long path-like reply trees (i.e., $S \approx D$ and large S). In addition, the distribution of D is similar between the model and data (Fig. 2b). However, model 3 is yet unsatisfactory for the following reasons. First, model 3 overestimates the probability of S at large S (Fig. 2a). Second, as shown in Fig. 6a, long path-like trees are much fewer in model 3 than in the empirical data for the entire range of d_1 examined in the figure. Third, the model does not produce sufficiently many large star-like trees (at small d_2 in Fig. 6b). Fourth, the model overestimates the fraction of large irregular trees (Fig. 6c, d).

4.3.2 Model 4

Empirically, Fig. 7b indicates that λ decreases on an average with the in-degree of the root tweet (i.e., k_t). Therefore, we extend model 3 by assuming that the distribution of λ depends on the in-degree of the tweet at the root of the reply tree. In the extended model, which we refer to as model 4, we draw the length of each segment from $P(\lambda|k_t)$. Then, the in-degree of the end node of each segment is drawn from $P(k_e)$ constructed from the empirical data, which is the same as in model 3 (Table 1).

The distributions of both S (Fig. 2a) and D (Fig. 2b) are close between model 4 and the empirical data. The joint distribution of S and D obtained from the model is shown in Fig. 3e. The fraction of long path-like trees for small d_1 is similar between the model and data (Fig. 6a). However, large star-like trees (see Fig. 6b where d_2 is small), and large irregular trees (Fig. 6c, d) are considerably fewer for the model than the empirical data.

4.3.3 Model 5

We consider a further extension of the model in which k_e for each end node of a segment is drawn from the empirically constructed conditional distribution $P(k_e|k_t)$ instead of the unconditional distribution $P(k_e)$ employed in models

3 and 4. It should be noted that the end node does not have to be that for the segment emanating from the root. We refer to the extended model as model 5. This extension of the model is empirically supported; the in-degree of the end node of a segment considerably depends on the in-degree of the tweet that initiates the reply tree (Fig. 7c).

Similarly to model 4, model 5 produces the distributions of S (Fig. 2a) and D (Fig. 2b) that are close to the empirical data. The joint distribution $P(S, D)$ for model 5 is shown in Fig. 3f. The fraction of long path-like trees with small d_1 (Fig. 6a) and that of large star-like trees for a range of d_2 (Fig. 6b) are not far from those for the empirical data. However, the model produces much less irregular trees than the empirical data (Fig. 6c, d).

4.4 Models with correlated segment lengths

4.4.1 Empirical evidence of correlated segment lengths

The models introduced so far are incapable of producing a realistic frequency of large irregular trees. Although large irregular trees are rare even in the empirical data (Fig. 6c, d), they are suggestive of mechanisms that generate an entire reply tree. If the λ values for the k_s segments starting from the same node are positively correlated, large irregular trees are expected to occur relatively easily. This is because a large λ value in one branch implies a relatively high probability of large λ values in other branches in the same reply tree. For example, if the root has in-degree 2, both of the two segments have $\lambda = 100$, and no further branching occurs, we obtain a large irregular tree with $S = 201$ and $D = 100$.

Denote by $\langle \lambda \rangle_s$ the average of λ over the k_s segments starting from the same node. If the λ values for the k_s segments are positively correlated, $\langle \lambda \rangle_s$ statistically fluctuates more than realizations of $\langle \lambda \rangle_s$ calculated on the basis of independent λ values as we assumed in models 3, 4, and 5. In the independent case, $\langle \lambda \rangle_s$ has standard deviation equal to $\sigma(\lambda)/\sqrt{k_s}$, where $\sigma(\lambda)$ is the standard deviation of λ calculated from $P(\lambda)$. It should be noted that the mean of $\langle \lambda \rangle_s$ is the same between the empirical and the independent cases, because we use the empirical $P(\lambda)$ to independently draw λ values for the k_s segments in models 3, 4, and 5.

The standard deviation of $\langle \lambda \rangle_s$ calculated from the empirical data and that calculated from $P(\lambda)$ under the independence assumption are plotted against k_s in Fig. 7d. The figure suggests that the fluctuation of $\langle \lambda \rangle_s$ is larger for the empirical data than under the independence assumption unless k_s is large. The amount of fluctuation is the same between the two cases when k_s is large. Therefore, λ observed in the empirical data may be positively correlated across segments sharing a start node.

4.4.2 Models 3, 4, 5 with copula

Motivated by the results shown in Sect. 4.4.1, we extend models 3, 4, and 5 (Sect. 4.3) to allow λ to be correlated among segments emanating from the same starting node as follows. For each start node, we use a k_s -dimensional multivariate normal distribution to generate k_s correlated variables denoted by (x_1, \dots, x_{k_s}) . We assume that each x_i ($1 \leq i \leq k_s$) is distributed according to the standard normal distribution (i.e., mean zero and standard deviation one) when marginalized. Then, we transform each x_i to λ_i , the value of λ for the i th segment, such that the marginal distribution of λ_i coincides with the empirical $P(\lambda)$.

To realize this goal, we generate (x_1, \dots, x_{k_s}) using the multivariate standard normal distribution with mean $(0 \dots 0)$ and the covariance matrix given by

$$\Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix}, \quad (1)$$

where ρ is the covariance of the random variables corresponding to x_i and x_j , which is common for all pairs of i and j ($i \neq j$). Because Σ is a positive semi-definite matrix and the eigenvalues of Σ are given by $1 - \rho$, which is $(k_s - 1)$ -fold, and $1 + (k_s - 1)\rho$, which is nondegenerate, we obtain $-1/(k_s - 1) \leq \rho \leq 1$. It should be emphasized that, regardless of the value of ρ , x_i ($1 \leq i \leq k_s$) obeys the standard normal distribution when marginalized. The numerical method with which we generate samples from the correlated multivariate normal distribution (Rachev 2003) is explained in Appendix 1.

Then, we map x_i onto $[0, 1]$ by the error function using

$$y_i = \int_{-\infty}^{x_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx. \quad (2)$$

Each y_i is uniformly distributed on $[0, 1]$, and different y_i 's are correlated with each other when $\rho \neq 0$. This method for generating a correlated random ensemble on a hypercube is called a Gaussian copula (Rachev 2003; Franke et al. 2008). Then, we transform y_i to λ_i by finding the unique value of λ_i that satisfies

$$\sum_{\lambda=1}^{\lambda_i-1} P(\lambda) \leq y_i \leq \sum_{\lambda=1}^{\lambda_i} P(\lambda). \quad (3)$$

Each λ_i obeys $P(\lambda)$ when marginalized, and the strength of the correlation between different λ_i 's is controlled by ρ . When $\rho = 0$, we reproduce the models considered in Sect. 4.3.

We numerically examine the copula variants of models 3, 4, and 5, which by definition employ distributions of λ .

We set $\rho = 0.7$, which we found to produce results relatively close to the empirical data. Figure 2c indicates that model 3 with copular overestimates the probability of S at large S , whereas models 4 and 5 with copula produce distributions of S close to the empirical one. All three models with copular produce distributions of D close to the empirical one (Fig. 2d). The joint distributions $P(S, D)$ for the three copula models are shown in Fig. 3g–i. The fraction of long path-like reply trees, that of large star-like reply trees, and that of irregular reply trees are shown in Fig. 6e–h. Similar to the distribution of S (Fig. 2c), the figures indicate that introduction of the copula improves models 4 and 5, but not model 3. Figure 6e indicates that the fraction of long path-like trees is similar among models 4 with copula, model 5 with copula, and the empirical data when d_1 is small, although significant discrepancies remain for large d_1 . Figure 6f indicates that the fraction of large star-like trees is close between model 5 with copula and the empirical data over the entire range of d_2 . It should be noted that model 4 with $\rho = 0$ and $\rho = 0.7$ produces much less star-like trees than the empirical data when d_2 is small. Figure 6g, h indicate that the fraction of large irregular trees is also similar among model 4 with copula, model 5 with copula, and the empirical data for the entire parameter range explored by our numerical simulations. We conclude that model 5 with $\rho = 0.7$ captures main statistical properties of reply trees observed empirically, despite a notable discrepancy in the frequency of long path-like trees when d_1 is not small (Fig. 6e).

5 Conclusions

We analyzed structure of reply trees observed in Twitter. We examined a suite of branching process models to capture properties of empirical data in terms of the frequency of long path-like reply trees, large star-like reply trees, and large irregular reply trees, which are typologies proposed in Cogan et al. (2012), as well as the distributions of the size and depth of reply tree. The Galton–Watson process and its correlated variant did not produce realistic statistics of reply tree. Our final model (i.e., model 5 with copula) assumed that the segment length (i.e., λ) and the degree of end nodes of segments depended on the in-degree of the tweet located at the root of the reply tree. These assumptions imply that the tweet at the root, whose first-order properties may be encoded in its in-degree, seems to be a strong determinant of the shape of the reply tree (Wang et al. 2011; Li et al. 2012; Gómez et al. 2013). The final model also assumed that λ was positively correlated among segments starting from the same node. This assumption is also in line with the idea that the in-degree of the tweet affects the entire topology of the reply tree for the

following reason. Owing to their contents, some tweets may tend to induce long segments in the reply trees rooted at them. Other tweets may tend to induce short segments. If this is the case, different segments in a reply tree would be positively correlated. For simplicity, in our models, we introduced positive correlation only to segments sharing the start node.

Copulas have been used for generating correlated networks (Gleeson 2008; Raschke et al. 2014). In these studies, two-dimensional copulas were used for defining the joint degree distribution of an adjacent pair of nodes. In contrast, the present study employed a Gaussian copula of a general dimension to produce correlated segments sharing a start node.

A serious limitation of the present development is that we have plugged the empirical distribution of the length of segments, e.g., $P(\lambda)$, directly into models 3, 4, and 5. Then, we focused on other structural properties of reply trees such as correlation between segments sharing the start node. However, the mechanisms governing such correlations are not clear. In addition, users seem not to care about the length of segments when deciding whether or not to reply to other posts. Branching process models have also been criticized of not being able to explain other aspects of networks of posts (Kumar et al. 2010; Wang et al. 2011).

An alternative, agent-based approach is growing network models, in which a node with out-degree one joins an existing tree according to a certain attachment rule. This approach, which has been used for modeling networks of posts (Götz et al. 2009; Kumar et al. 2010; Li et al. 2012; Wang et al. 2012; Gómez et al. 2013; Gleeson et al. 2014), may be also useful for understanding the current data set.

Acknowledgments N.M. acknowledges the support provided through CREST, JST. M.T. was partially supported by JSPS KAKENHI Grant Number 25280111.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix 1: Analysis of the first half of the data

To examine the robustness of the results shown in the main text, we analyzed the reply trees whose tweet located at the root of the tree was posted in the first half of the observation period, i.e., between December 1, 2011 and the noon

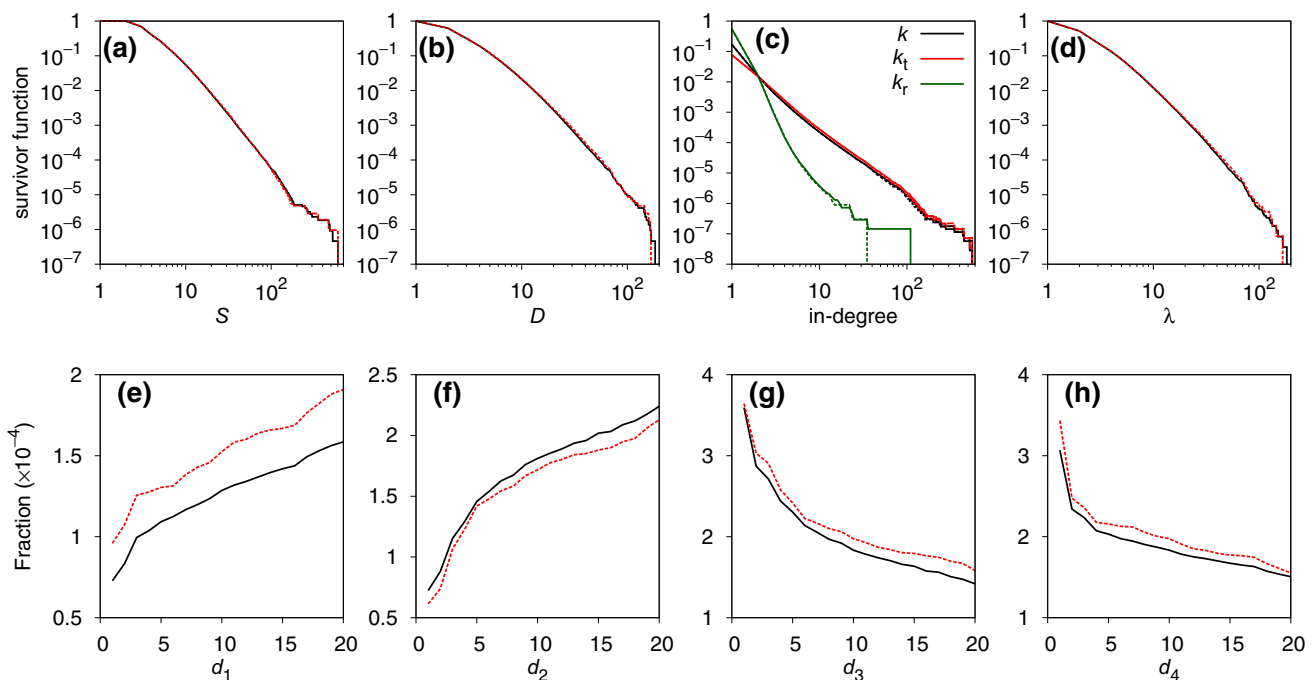


Fig. 8 Comparison between the results obtained from all reply trees (solid lines) and those obtained from the reply trees starting in the first half of the observation period (dashed lines). The solid lines are identical to those shown in the previous figures. **a** Survivor function of the size of the reply tree. **b** Survivor function of the depth of the reply tree. **c** Survivor functions of the in-degree for all nodes (i.e., k),

that restricted to tweets (i.e., k_t), and that restricted to replies (i.e., k_r). **d** Survivor function of the length of segment. **e** Fraction of long path-like trees. **f** Fraction of large star-like trees. **g** Fraction of large irregular trees when $d_3 = 10$. **h** Fraction of large irregular trees when $d_4 = 10$

of December 5, 2011. The data contained 4,375,861 nodes, of which there were 1,042,721 tweets and 3,333,140 replies. In Fig. 8, the dashed lines represent the results for the first half of the data, and the solid lines represent those for the entire data. The latter results are equivalent to those shown in Figs. 2, 4, 5, and 6 by the same types of line. The results are similar between the reduced and full data sets.

Appendix 2: Sampling from the correlated multivariate normal distribution

We generated random variables obeying the multivariate normal distribution with mean zero and the covariance matrix given by Eq. (1) as follows. It holds that

$$\Sigma = AA^{\top}, \quad (4)$$

where

$$A = \begin{pmatrix} a_1 & a_2 & \cdots & a_2 \\ a_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_2 \\ a_2 & \cdots & a_2 & a_1 \end{pmatrix}, \quad (5)$$

$$a_1 = \frac{(k_s - 1)\sqrt{1 - \rho} + \sqrt{(k_s - 1)\rho + 1}}{k_s}, \quad (6)$$

$$a_2 = \frac{-\sqrt{1 - \rho} + \sqrt{(k_s - 1)\rho + 1}}{k_s}, \quad (7)$$

and \top denotes the transposition. Therefore, by setting

$$\begin{pmatrix} x_1 \\ \vdots \\ x_{k_s} \end{pmatrix} = A \begin{pmatrix} z_1 \\ \vdots \\ z_{k_s} \end{pmatrix}, \quad (8)$$

where z_i ($1 \leq i \leq k_s$) is drawn from the standard normal distribution independent for different i , we obtain the desired $(x_1 \dots x_{k_s})$.

References

- Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the 4th ACM international conference on web search and data mining (WSDM '11). ACM press, New York, pp 65–74
- Banões RA, Borge-Holthoefer J, Moreno Y (2013) The role of hidden influentials in the diffusion of online information cascades. *EPJ Data Sci* 2(1):6
- Bliss CA, Kloumann IM, Harris KD, Danforth CM, Dodds PS (2012) Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *J Comput Sci* 3(5):388–397
- Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. *J Comput Sci* 2(1):1–8
- Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, Settle JE, Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–298
- Cha M, Mislove A, Gummadi KP (2009) A measurement-driven analysis of information propagation in the Flickr social network. In: Proceedings of the 18th international conference on world wide web (WWW '09). ACM press, New York, pp 721–730
- Cha M, Haddadi H, Benevenuto F, Gummadi KP (2010) Measuring user influence in Twitter: the million follower fallacy. In: Proceedings of the 4th international AAAI conference on weblogs and social media (ICWSM '10). AAAI Press, Palo Alto, pp 10–17
- Cheng J, Adamic LA, Dow PA, Kleinberg J, Leskovec J (2014) Can cascades be predicted? In: Proceedings of the 23rd international conference on world wide web (WWW '14). ACM Press, New York, pp 925–935
- Cogan P, Andrews M, Bradonjic M, Kennedy WS, Sala A, Tucci G (2012) Reconstruction and analysis of Twitter conversation graphs. In: Proceedings of the 1st ACM international workshop on hot topics on interdisciplinary social networks research (HotSocial '12). ACM Press, New York, pp 25–31
- Conover MD, Davis C, Ferrara E, McKelvey K, Menczer F, Flammini A (2013) The geospatial characteristics of a social movement communication network. *PLoS One* 8(3):e55957
- Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: hedonometrics and twitter. *PLoS One* 6(12):e26752
- Easley D, Kleinberg J (2010) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, Cambridge
- Fortunato S, Castellano C (2007) Scaling and universality in proportional elections. *Phys Rev Lett* 99(13):138701
- Franke J, Härdle WK, Hafner CM (2008) Statistics of financial markets: an introduction, 2nd edn. Springer, Berlin
- Gladwell M (2000) The tipping point: how little things can make a big difference. Little, Brown and Company, New York
- Gleeson JP (2008) Cascades on correlated and modular random networks. *Phys Rev E* 77(4):046117
- Gleeson JP, Ward JA, O'Sullivan KP, Lee WT (2014) Competition-induced criticality in a model of meme popularity. *Phys Rev Lett* 112(4):048701
- Golub B, Jackson MO (2010) Using selection bias to explain the observed structure of Internet diffusions. *Proc Natl Acad Sci USA* 107(24):10833–10836
- Gómez V, Kaltenbrunner A, López V (2008) Statistical analysis of the social network and discussion threads in Slashdot. In: Proceedings of the 17th international conference on world wide web (WWW '08). ACM Press, New York, pp 645–654
- Gómez V, Kappen HJ, Kaltenbrunner A (2011) Modeling the structure and evolution of discussion cascades. In: Proceedings of the 22nd ACM conference on Hypertext and hypermedia (HT '11). ACM Press, New York, pp 181–190
- Gómez V, Kappen HJ, Litvak N, Kaltenbrunner A (2013) A likelihood-based framework for the analysis of discussion threads. *World Wide Web* 16(5–6):645–675
- Gonçalves B, Perra N, Vespignani A (2011) Modeling users' activity on Twitter networks: validation of Dunbar's number. *PLoS One* 6(8):e22656
- Götz M, Leskovec J, McGlohon M, Faloutsos C (2009) Modeling blog dynamics. In: Proceedings of the 3rd international AAAI conference on weblogs and social media (ICWSM '09). AAAI Press, Palo Alto, pp 26–33
- Harris TE (1963) The theory of branching processes. Springer, Berlin
- Iribarren JL, Moro E (2009) Impact of human activity patterns on the dynamics of information diffusion. *Phys Rev Lett* 103(3):038702

- Iribarren JL, Moro E (2011) Branching dynamics of viral information spreading. *Phys Rev E* 84(4):046116
- Jo H-H, Perotti JI, Kaski K, Kertész J (2014) Analytically solvable model of spreading dynamics with non-Poissonian processes. *Phys Rev X* 4(1):011041
- Kimmel M, Axelrod DE (2002) *Branching processes in biology*. Springer, New York
- Kumar R, Mahdian M, McGlohon M (2010) Dynamics of conversations. In: *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '10)*. ACM Press, New York, pp 553–562
- Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: *Proceedings of the 19th international conference on world wide web (WWW '10)*. ACM Press, New York, pp 591–600
- Leskovec J, Singh A, Kleinberg J (2006) Patterns of influence in a recommendation network. In: *Advances in knowledge discovery and data mining*, Springer, Berlin. *Lecture notes in computer science*, vol 3918, pp 380–389
- Leskovec J, Adamic LA, Huberman BA (2007) The dynamics of viral marketing. *ACM Trans Web* 1(1):5
- Leskovec J, McGlohon M, Faloutsos C, Glance N, Hurst M (2007) Patterns of cascading behavior in large blog graphs. In: *Proceedings of the 7th SIAM international conference on data mining (SDM '07)*, SIAM, pp 551–556
- Li H, Liu J, Xu K, Wen S (2012) Understanding video propagation in online social networks. In: *Proceedings of IEEE 20th international workshop on quality of service (IWQoS)*, IEEE, pp 1–9
- Liben-Nowell D, Kleinberg J (2008) Tracing information flow on a global scale using Internet chain-letter data. *Proc Natl Acad Sci USA* 105(12):4633–4638
- Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I, Boyd D (2011) The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions. *Int J Commun* 5:1375–1405
- Lovejoy WS, Loch CH (2003) Minimal and maximal characteristic path lengths in connected sociomatrices. *Soc Netw* 25(4):333–347
- McGlohon M, Leskovec J, Faloutsos C, Hurst M, Glance N (2007) Finding patterns in blog shapes and blog evolution. In: *Proceedings of the international AAAI conference on weblogs and social media (ICWSM '07)*. AAAI Press, Palo Alto
- Olofsson P (1996) Branching processes with local dependencies. *Ann Appl Probab* 6(1):238–268
- Rachev ST (ed) (2003) *Handbook of heavy tailed distributions in finance*, *handbooks in finance*, vol 1. Elsevier Science B. V., Amsterdam
- Raschke M, Schläpfer M, Trantopoulos K (2014) Copula-based modeling of degree-correlated networks. *J Stat Mech* 2014:P02019
- Rogers EM (2003) *Diffusion of innovations*, 5th edn. Free Press, New York
- Sano Y, Yamada K, Watanabe H, Takayasu H, Takayasu M (2013) Empirical analysis of collective human behavior for extraordinary events in the blogosphere. *Phys Rev E* 87(1):012805
- Sasahara K, Hirata Y, Toyoda M, Kitsuregawa M, Aihara K (2013) Quantifying collective attention from tweet stream. *PLoS One* 8(4):e61823
- Sousa D, Sarmento L, Rodrigues EM (2010) Characterization of the twitter @replies network: are user ties social or topical? In: *Proceedings of the 2nd international workshop on search and mining user-generated contents (SMUC '10)*, ACM Press, New York, pp 63–70
- Sun E, Rosenn I, Marlow CA, Lento TM (2009) Gesundheit! Modeling contagion through Facebook news feed. In: *Proceedings of the 3rd international AAAI conference on weblogs and social media (ICWSM '09)*, AAAI Press, Palo Alto, pp 146–153
- Thelwall M, Haustein S, Larivière V, Sugimoto CR (2013) Do altmetrics work? Twitter and ten other social web services. *PLoS One* 8(5):e64841
- Vazquez A, Rácz B, Lukács A, Barabási A-L (2007) Impact of non-Poissonian activity patterns on spreading processes. *Phys Rev Lett* 98(15):158702
- Wang D, Wen Z, Tong H, Lin C-Y, Song C, Barabási A-L (2011) Information spreading in context. In: *Proceedings of the 20th international conference on world wide web (WWW '11)*. ACM Press, New York, pp 735–744
- Wang C, Ye M, Huberman BA (2012) From user comments to on-line conversations. In: *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '12)*. ACM Press, New York, pp 244–252
- Weng J, Lim E, Jiang J, He Q (2010) TwitterRank: finding topic-sensitive influential twitterers. In: *Proceedings of the 3rd ACM international conference on web search and data mining (WSDM '10)*. ACM Press, New York, pp 261–270